JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

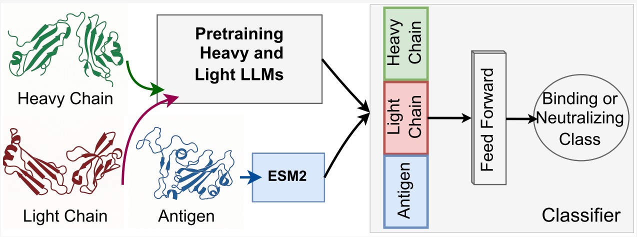
pubs.acs.org/jcim Article

Predicting Antibody−Antigen Interactions with Structure-Aware LLMs: Insights from SARS-CoV-2 Variants

Faisal Bin Ashraf, Vinz Angelo Madrigal, and Stefano Lonardi*

Cite This: https://doi.org/10.1021/acs.jcim.5c00973 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Predicting antibody−antigen interactions is a critical step in developing new therapeutics to defend against viral infections. However, measuring the extent of these interactions *in vitro* is costly and time-consuming. With the increased availability of experimental data, predictive methods using machine learning, particularly large language models (LLMs), have emerged as a powerful alternative to wet lab experiments. Here we focus on antibodies targeting SARS-CoV-2 variants, given the abundance of data on this highly contagious virus and the impact of COVID-19 on human life. The objective of this work is to predict the binding and the neutralizing properties of SARS-CoV-2 antibodies. While there are many studies on predicting binding, to the best of our knowledge, we are the first to address the problem of predicting the neutralizing properties of SARS-CoV-2 antibodies. Here we propose a new classifier that combines LLMs with structural information. Extensive experimental results show our method (i) achieves high prediction accuracy (especially for closely related antigen variants) and (ii) outperforms other classifiers in the literature on the prediction of antibody−antigen binding.

■ INTRODUCTION

The interaction between antibodies and antigens is the fundamental mechanism of action of the immune system, where specialized proteins called *antibodies* recognize and bind to specific molecular targets known as *antigens*.[1,2] This molecular recognition process is crucial for the organism's defense against pathogens, enabling the precise identification and neutralization of potential threats.[3] Most interactions occur in the *complementarity-determining regions* (CDR) of antibodies, which are designed to target specific antigens with high specificity and precision.[4] The significance of these interactions extends far beyond basic immunological research.[5] By elucidating the complex "molecular rules" of how antibodies bind to antigens, scientists can develop more effective therapeutic interventions, design vaccines, and diagnostic tools.[6] These interactions are not just docking events, but complex molecular processes involving intricate structural arrangements, electrostatic interactions, and sophisticated recognition mechanisms that allow the immune system to distinguish between self-and nonself molecules with high accuracy.[7,8]

Recent technical advances in machine learning (ML) have revolutionized the prediction and understanding of antibody−antigen interactions, addressing critical challenges in computational immunology.[9] Traditional laboratory experimental methods that can detect and measure these interactions are time-consuming and resource-intensive, making computational approaches increasingly essential.[10] ML models, particularly deep learning architectures like graph neural networks and transformer-based models, can now predict antibody−antigen binding affinities, epitope locations, and potential interaction sites (see, e.g.,[11−17]). These computational approaches not only promise to accelerate drug discovery and vaccine design but also have the potential to provide insights into the complex molecular recognition mechanisms that are challenging to explain through conventional experimental techniques.[18] By leveraging the availability of large-scale structural databases and advanced machine learning methods, researchers can now deploy predictive models that can reduce screening time, detect antibody−antigen interactions for novel pathogens, and design more targeted therapeutic interventions.[10]

Recent innovations in large language models (LLMs) have further transformed the landscape of antibody−antigen interaction prediction tasks.[19] Unlike traditional structure-based methods that require extensive computational resources and detailed structural information, sequence-based approaches leverage the power of LLMs to extract meaningful patterns directly from amino acid sequences.[20] Antibody-specific LLMs can now predict interactions, binding affinities, and functional characteristics by learning from vast databases of antibody sequences.[21] For instance, several studies have explored the use of LLMs for antibody design and epitope

Received: April 30, 2025
Revised: August 4, 2025
Accepted: August 6, 2025

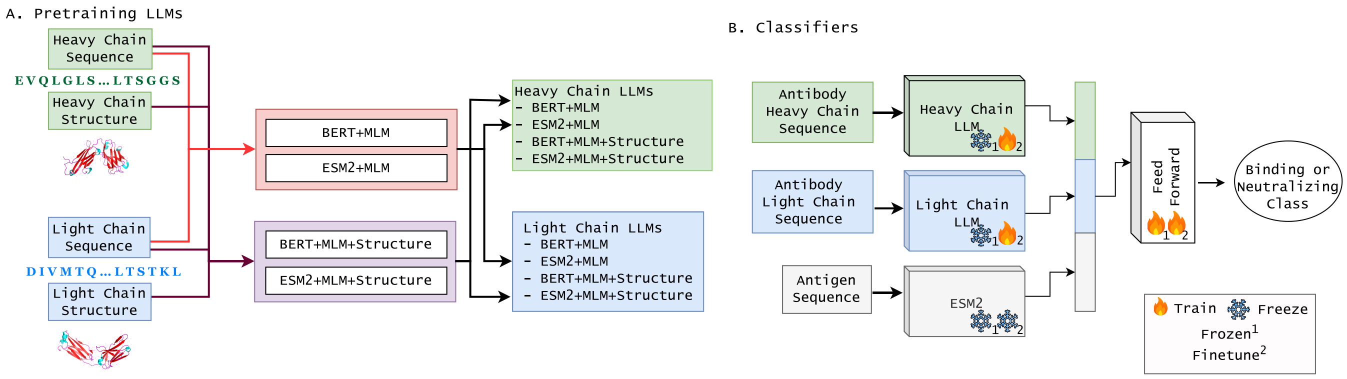


Figure 1. Overview of the study. (A) Several LLMs have been trained in this study on the heavy and light chain sequences of antibodies; BERT + MLM and ESM2 + MLM were trained with sequences only; BERT + MLM + Structure and ESM2 + MLM + Structure were trained using both sequence and structure; MLM = Masked Language Modeling (B). The proposed architecture for predicting binding and neutralizing properties of antibodies; it combines one heavy chain LLM, one light chain LLM (both pretrained in step A), and one ESM2 LLM.

prediction for SARS-CoV-2.[22,23] Approaches based on LLMs offer several critical advantages: reduced computational cost, ability to handle proteins with unknown structures, faster screening of potential protein−protein interactions, and the capacity to generalize across diverse protein families.[19,24] Moreover, sequence-based models can capture subtle evolutionary and contextual relationships, providing insights into protein−protein interactions that go beyond simple geometric matching.[25]

While sequence-based models offer significant advantages, they inherently miss the critical structural context that defines antibody−antigen interactions.[26] Structural information provides crucial insights into molecular recognition that cannot be fully captured by sequence data alone.[27] The three-dimensional arrangement of amino acids and their precise interactions has a fundamental role in determining antibody−antigen binding specificity and affinity.[28] Consequently, integrating structural data within LLMs has great potential to advance the field of computational immunology.[29] Recent developments in AlphaFold and other AI-driven structure prediction technologies have enabled the generation of high-accuracy protein structures from sequences, bridging the gap between sequence and structural information.[30] By combining synergistically the pattern recognition capabilities of LLMs with the structural insights provided by AlphaFold, more comprehensive and accurate predictive models of protein−protein interactions have been developed.[31] This hybrid approach allows for a more holistic understanding of molecular recognition, capturing both the evolutionary context encoded in sequences and the critical spatial arrangements that determine binding specificity.[32] For instance, ESMFold offers a practical and scalable alternative to AlphaFold for protein structure prediction.[33] ESMFold is more computationally efficient than AlphaFold and does not require users to provide a multiple sequence alignment. Extensive benchmarking showed that ESMFold achieves comparable performance to AlphaFold on some classes of proteins (e.g., antibodies), and it can provide global fold and framework architectures which are most relevant for auxiliary structure-based modeling.[34] The idea of combining sequence-based modeling with evolutionary, structural or functional information has been explored in the literature (e.g., AlphaFold Evoformer,[27] MSA Transformer[35]). Some models focus on training separate encoders for sequence and structure and then combine them to predict the properties (see, e.g.,[36,37]). Other approaches use a structure-aware vocabulary[38] or remote homology[39] to force the model to learn the structural features.

The problem of predicting the interaction of antibodies against the same class of target antigen remains, however, relatively unexplored. Most studies in the literature focus on antibody−antigen interactions across various pathogens, where significant differences in antigen sequence and structure exist. However, when antigens differ by only a few amino acid substitutions, existing predictive models often struggle to accurately capture changes in interactions. It has been shown that minor mutations in the antigen sequence can significantly influence antibody binding and neutralization, highlighting the challenges in modeling such subtle yet impactful changes.[40]

In this study, we evaluate the ability of a LLM to predict the binding specificity and neutralizing properties of antibodies targeting various variants of SARS-CoV-2. The key finding of our study is that incorporating structural information allows the LLM to more effectively capture the binding and neutralizing properties of antibodies against variants of the same pathogen.

■ MATERIALS AND METHODS

First, we pretrained two language models individually for the heavy and light chains using the *masked language modeling* (MLM) objective. The language models were trained with and without structural information. For heavy and light chains, we explored four variants, namely BERT-based MLM, ESM2-based MLM, and their respective structure-augmented versions. After pretraining, we developed classifiers that combine the embeddings for the heavy chain, the embeddings for the light chain, and antigen sequences to predict binding or neutralization classes.

In the classifier architecture, heavy chain and light chain models are either frozen or fine-tuned, while the antigen sequence is encoded using a frozen ESM2 model. The embeddings from each branch are concatenated and passed through a feedforward network for final classification. This two-stage design—pretraining followed by classification—enables a systematic assessment of how structure augmentation and fine-tuning impact downstream performance. Figure 1 shows the outline of our method.

Data Set. Our training data set was derived by the CovAbDab database (Feb 2024),[41] which contains a comprehensive collection of annotated antibodies known to interact with various SARS-CoV-2 variants, including the original Wuhan strain (hereafter WT), as well as β, Delta, and Omicron variants. CovAbDab provides detailed annotations for each antibody, including amino acid sequences for heavy and light chains, the host species, binding specificity, and neutralization properties. Additionally, structural data, precise epitope mappings, and germline gene information are often available.

In this study, the *antigen* of interest is the spike protein of SARS-CoV-2. However, CovAbDab does not contain annotations for SARS-CoV-2 variants; thus, we manually obtained these variant sequences from the Protein Data Bank (PDB) hosted at RCSB.[42] The sequence variations in the spike protein for β, Delta, and Omicron variants relative to the WT sequence include amino acid substitutions, insertions, and deletions. Figure 2 illustrates the positions of these variations for all the SARS-CoV-2 variants. Observe that (1) the most common variations are substitutions, and (2) most substitutions occur prominently within the receptor-binding domain (RBD, positions 305–534), (3) insertions and deletions predominantly appear in the N-terminal domain (positions 13–304). The Levenshtein (edit) distance between the WT and other SARS-CoV-2 variants ranges from 14 to 57 amino acids (data not shown). Given that the total length of the spike protein sequence of WT is 1208 residues, these distances represent approximately 3% sequence divergence, highlighting the high similarity among the variant antigen sequences.

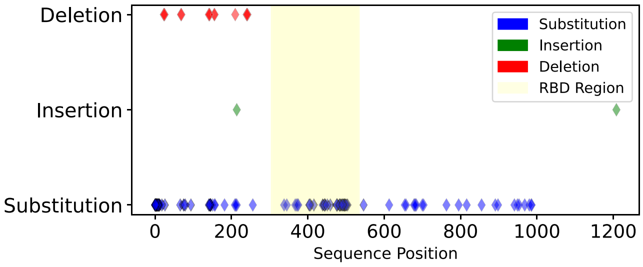


Figure 2. Positions of amino acid substitutions, insertions and deletions of SARS-CoV-2 variants with respect to the original Wuhan strain (WT).

We processed the CovAbDab data set to determine the list of antibody—antigen binding pairs and the list of antibody—antigen neutralizing pairs. Since every antibody—antigen interaction is not equally represented in CovAbDab, we restricted the analysis to variants with at least 1000 interacting antibodies to ensure robust statistical support and reliability in our results. This subset of CovAbDab contained 42,091 antibody—antigen pairs. Table 1 summarizes the number of antibody—antigen pairs for which the binding and neutralizing properties are known. Note that neutralizing antibodies represent a subset of antibodies that bind to their respective target antigens. It is not possible for an antibody to be neutralizing but not binding. Observe that SARS-CoV-2 WT has the highest number of antibodies in the data set. This likely reflects annotation bias, as the original Wuhan strain emerged the earliest, prompting extensive initial research and antibody characterization. Newer variants have fewer annotated antibodies due to comparatively limited studies and experimental data. We also created a *combined label* that captures three interaction modalities, namely (i) binding and neutralizing, (ii) binding but not neutralizing, and (iii) neither binding nor neutralizing.

Next, we determined which pairs of amino acids in the antibodies are in close proximity because those pairs are more likely to form contacts. According to the guidelines used in the Critical Assessment of protein Structure Prediction (CASP) competition, two amino acids are considered *in contact* when the distance between their α-carbon atoms is less than 8° A.[43] The subset of CovAbDab used in this study contained 10,386 distinct antibodies, of which only 2237 had an experimentally determined 3D structure available in the Protein Data Bank (PDB). To address this problem, we predicted the structures of the missing antibodies using ESMFold.[33] Then, we calculated all pairwise distances between the α carbon C_α atom on all residue—residue pairs. Given the 3D structure of an antibody S of length |S|, we generated a |S| × |S| binary matrix contact map, where element $(i, j) = 1$ if the C_α atom of residues $S[i]$ and $S[j]$ were closer than 8° A, $(i, j) = 0$ otherwise. This process was carried out on the heavy chain and the light chain of all the antibodies in the data set.

Pretraining LLMs on Sequence and Structure. Protein Language Models (PLMs) are analogous to LLMs for natural language processing. PLMs are trained on large protein sequence data sets to capture evolutionary and structural information. They facilitate predicting protein interactions,

Table 1. Number of Antibodies Binding or Neutralizing the 14 SARS-CoV-2 Variants Used in this Study

variant	binding	not binding	neutralizing	not neutralizing
sars-cov-1	1723	1578	472	2829
sars-cov2-wt	8960	958	4425	5493
sars-cov2-β	827	123	419	531
sars-cov2-delta	979	221	510	690
sars-cov2-omicron-ba1	2288	1905	1903	2290
sars-cov2-omicron-ba1.1	790	696	775	711
sars-cov2-omicron-ba2	2292	1466	1724	2034
sars-cov2-omicron-ba2.12.1	812	848	773	887
sars-cov2-omicron-ba2.13	696	708	674	730
sars-cov2-omicron-ba2.75	1436	1896	1411	1921
sars-cov2-omicron-ba3	756	746	706	796
sars-cov2-omicron-ba4	607	891	567	931
sars-cov2-omicron-ba5	1786	2933	1655	3064
sars-cov2-omicron-xbb1	502	1786	485	2685
total	**24,454**	**17,637**	**16,499**	**25,592**

structures, and functional properties by leveraging patterns learned from sequence data. PLMs are typically trained using a *Masked Language Modeling* (MLM) objective to learn the "language of proteins" at the sequence level by minimizing the loss $\mathcal{L}_{\text{MLM}}$, defined as follows

$$\mathcal{L}_{\text{MLM}} = -\frac{1}{|M|}\sum_{i\in M}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c})$$

where M is the set of masked positions in the sequence, C is the vocabulary size (i.e., the number of all the possible amino acids in any position), $y_{i,c}$ is the ground truth for amino acid c in position i, and $\hat{y}_{i,c}$ is the predicted probability for amino acid c in position i.

It is well-known that the function of proteins is determined by their 3D structure. General PLMs are expected to learn protein functions from the primary sequences, which limits the predictive ability for downstream tasks that depend on the 3D structure. Using a contacts-based loss (structural information) guides the LLM to learn spatial proximity patterns between residues, inherently capturing structural and functional properties. This structural awareness can directly enhance the capability of the model to predict antibody binding and neutralizing properties, since these properties strongly depend on precise molecular interactions and 3D conformations. In this study, we incorporated the contact information between all pairs of amino acids in the antibodies as a training objective as follows

$$\mathcal{L}_{\text{contacts}} = -\frac{1}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{L} M_{i,j}\log(\hat{M}_{i,j})$$

where L is the length of the sequence, $M_{i,j}$ is the ground truth for the structural contact between amino acid i and j, and $\hat{M}_{i,j}$ is the predicted contact between amino acid i and j.

We used the following objective function to train our structure-aware LLMs.

$$\mathcal{L}_{\text{MLM+structure}} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{contacts}}$$

Figure 1A illustrates the pretraining step for all the LLMs used in this study. The input to each LLM was (1) the sequence for the heavy chain of the antibody, (2) the sequence for the light chain of the antibody, and (3) optionally, the contact map for the antibody obtained from its 3D structure. Input sequences were processed through BERT (Bidirectional Encoder Representations from Transformers)[44] or ESM2,[45] using the loss function $\mathcal{L}_{\text{MLM}}$. BERT is a popular transformer-based architecture for the analysis of biological sequences such as DNA and proteins.[46,47] Our *BERT + MLM* is a 12-layer architecture with eight attention heads that we trained from scratch using exclusively the sequences for the heavy chains and light chains. ESM2 is a transformer-based PLM that was trained on all the protein sequences from the UniRef database.[45] ESM2 is a state-of-the-art protein language model that provides latent space embeddings for downstream tasks such as stability prediction, function annotation, and viral fitness modeling.[48−50] We started from a pretrained ESM2, then fine-tuned it on our antibodies, which resulted in the *ESM2 + MLM* models for heavy and light chain sequences. For the structure-aware LLMs, we developed *BERT + MLM + Structure* and *ESM2 + MLM + Structure* for heavy and light sequences, using the objective function $\mathcal{L}_{\text{MLM+structure}}$. We trained BERT from scratch to assess the ability of an antibody-specific language model trained without prior protein knowledge. The comparison of ESM2 with a model trained from scratch enabled us to evaluate whether it is advantageous to incorporate general protein knowledge for antibody-specific modeling. We compared the performance of all these antibody-specific LLMs with a basic pretrained ESM2 to carry out the three prediction tasks.

The pretraining of these LLMs was carried out on an NVIDIA A100 GPU server, utilizing a batch size of 64 and a total of 30 epochs. Each epoch consisted of 123 batches. The input sequences were divided into 80% training and 20% test sets, with 15% of the training set reserved for validation. The learning rate was 5×10^{-5}, and gradient clipping was applied with a max gradient norm of 1.0. A masking rate of 0.3 was applied to the input sequences.

Training to Predict Binding and Neutralizing Properties. The final architecture for predicting antibody−antigen binding or neutralizing is shown in Figure 1B. It is composed of two pretrained language models that independently encode the heavy chain sequence (VH) and the light chain sequence (VL) of antibodies, combined with one ESM2 pretrained model for the sequence of the specific antigen. The concatenated embeddings produced by the three LLMs were processed through one classification layer, which mapped the combined embedding vector of dimension (2 × antibody embedding dimension + antigen embedding dimension) to single scalar output. These scalars were transformed using a sigmoid function to yield a probability representing the predicted likelihood of a binding or neutralizing event. We selected the smallest ESM-2 model (*t6_8M_UR50D*) to encode antigen sequences ($n = 14$), as training a larger model on such a limited data set would be unreasonable. As shown in Supporting Figure S1, the model produced distinct embeddings for each of the antigen sequences despite its compact size. The cross-entropy loss function $\mathcal{L}_{\text{CE}}$ used to train the classifiers is defined as follows

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c})$$

where N is the number of samples, C is the number of classes, $y_{i,c}$ is the class label, and $\hat{y}_{i,c}$ is the predicted probability for the ith sample and class c.

To establish a baseline for comparison, we utilized a basic architecture (hereafter referred to as *ESM2*) pretrained on the UniRef database. This baseline employs separate pretrained ESM2 models to independently represent the heavy chain, the light chain, and the antigen sequences. Additionally, we evaluated and compared four antibody-specific classifiers, namely *BERT + MLM*, *BERT + MLM + Structure*, *ESM2 + MLM*, and *ESM2 + MLM + Structure*. Each classifier was built upon one of our pretrained antibody-specific language models (LLMs), augmented with a feed-forward layer to classify antibodies according to their binding or neutralizing properties. These classifiers were chosen to assess how antibody-specific pretraining and inclusion of structural information impact the prediction of antibody binding and neutralizing properties compared to a general protein baseline. We used two model configurations for training these classifiers, namely

1. Frozen Configuration: In this case, both the heavy chain LLM and light chain LLM were frozen. Only the feed-forward layer was trained.

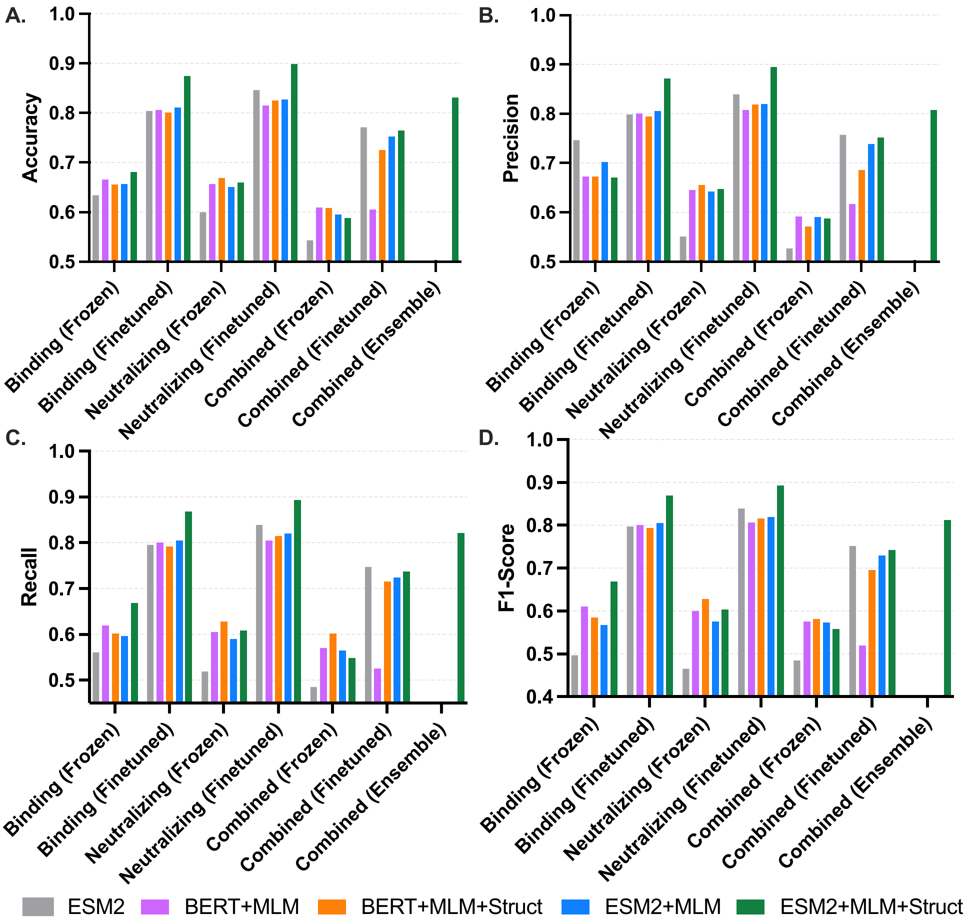


Figure 3. Performance evaluation of the five classifiers listed in the legend on all three predictive tasks (binding, neutralizing and combined); in the frozen configuration both the heavy chain LLM and light chain LLM were frozen and only the feed-forward layer was trained; in the finetuned configuration, the entire architecture was fine-tuned, with the exception of the LLM encoding the antigen (which was kept frozen); A: accuracy, B: precision, C: recall, and D: F1-Score (numerical values are available in Supporting Information).

2. Finetuned Configuration: In this case, the entire architecture was fine-tuned, with the exception of the LLM encoding the antigen (which was kept frozen).

The two configurations (Frozen and Finetuned) were chosen to evaluate the trade-off between computational efficiency and model adaptability. The Frozen configuration offers faster training and prevents overfitting but may limit model adaptability, whereas the Finetuned configuration enhances model flexibility at the cost of increased computational resources and potential overfitting. An ensemble of two pretrained models using the *ESM2* + *MLM* + *Structure* architecture, was also employed to separately predict the binding and neutralizing properties of antibody−antigen pairs. Based on these predictions, combined labels were assigned to each pair according to the rules described in the Data Set section.

The training data set for this step consisted of (1) the heavy chain sequence, (2) the light chain sequence, (3) the antigen sequence, and (4) the binding/neutralizing labels. Antibody sequences were tokenized using the same tokenizer used in the pretraining phase. Antigen embeddings were precomputed. Training was carried out using the binary cross-entropy with logits loss function, which is well-suited for binary classification tasks. For the prediction of the combined label (multiclass), the cross-entropy loss function was used. The model was optimized using the Adam optimizer with a learning rate of 1 × 10^{-5}, providing stability and efficient convergence. The input sequences were divided into 80% training and 20% test sets, with 15% of the training set reserved for validation. Each model was trained for 30 epochs on an NVIDIA A100 GPU. The model checkpoint with the highest validation accuracy was saved to avoid overfitting and to preserve the best-performing parameters.

■ RESULTS

Structure-Aware LLMs Outperform Sequence-Only LLMs. We evaluated the classifiers on their ability to predict binding, neutralizing, and combined labels for held-out test samples, under both frozen and finetuned configurations. Figure 3 summarizes the classification performance in terms of accuracy, precision, recall, and F1-score. Observe that the *ESM2* + *MLM* + *Structure* classifier achieved the best results for binding classification, with Accuracy, Precision, Recall, and F1-score reaching 0.8740, 0.8712, 0.8684, and 0.8697 respectively. Similar observations can be made for the neutralizing classification. Also, observe that finetuning consistently improved the performance of the classifiers across all tasks. Without finetuning, *ESM2* + *MLM* + *Structure* outperformed others, with the highest F1-score across tasks. *BERT* + *MLM* and *BERT* + *MLM* + *Structure* had modest performance, with significant gains after finetuning. However, it is clear from these experimental results that structure-aware LLMs (i.e.,

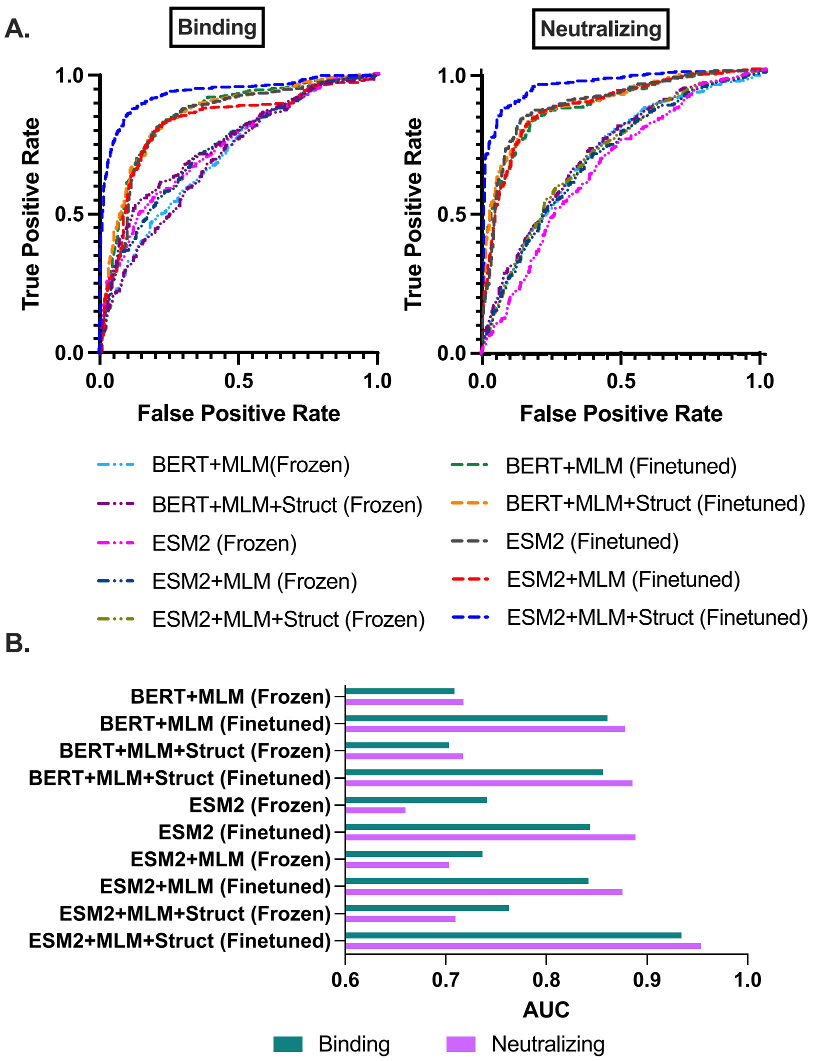


Figure 4. (A) ROC curve for binding and neutralizing classification tasks for all the classifiers in this study in both frozen and finetuned configurations; (B) AUC values for both classification tasks for all the classifiers in this study in both frozen and finetuned configurations (numerical values are available in Supporting Information).

BERT + MLM + Structure, ESM2 + MLM + Structure) performed better across all tasks than sequence-based LLMs (i.e., *ESM2, BERT + MLM, ESM2 + MLM*).

On the multiclass prediction of the combined class labels, the *ESM2* classifier performed very well when it was finetuned. Otherwise, our trained LLMs have better performance. However, when we used an ensemble of the finetuned *ESM2 + MLM + Structure* binding classifier and the *ESM2 + MLM + Structure* neutralizing classifier, we obtained the best performance. The ensemble of these two classifiers achieved an accuracy of 0.8310 and an F1-score of 0.8126 (as shown in Figure 3).

Figure 4A shows the Receiver Operating Characteristic (ROC) curves comparing the performance of five classifiers on binding and neutralizing prediction tasks, evaluated on held-out test samples in both frozen and fine-tuned configurations. Figure 4B shows the corresponding area under the curve (AUC) scores. Across both tasks, finetuned classifiers consistently outperformed their frozen counterparts. The finetuned *ESM2 + MLM + Structure* classifier achieved the highest AUC in both tasks (0.9342 for binding and 0.9538 for neutralizing) and the highest ROC curve.

We further evaluated the performance of these classifiers across variants of the target antigen. Figure 5 shows the classification F1-score for various choices of the antigen. Observe that the *ESM2 + MLM + Structure* classifier performed well on all variants except (i) SARS-Cov-1, which has low sequence similarity to other variants,[51] and (ii) SARS-Cov2-omicron-xbb1, which has highly distinguished functionalities compared to other variants.[52,53]

Finally, we compared the *ESM2 + MLM + Structure* classifiers against other deep learning and LLM-based binding prediction methods in the literature, namely DeepAIR,[54] Dynamic Masking LLM,[55] Ens-Grad,[56] ESM-F,[23] AntiBERTa,[23] AbMap,[57] and A2Binder.[23] DeepAIR uses a feature-encoding backbone and multiple task-specific prediction layers, and it was trained on curated experimental data sets. Other methods include Dynamic Masking LLM, which employs preferential masking of the CDR3 regions rather than uniform masking across the entire sequence and was pretrained on a large data set of natively paired antibody sequences. Ens-Grad utilizes a CNN-based architecture trained on sequences derived from phage display experiments. ESM-F and AntiBERTa are LLMs trained specifically on antibody

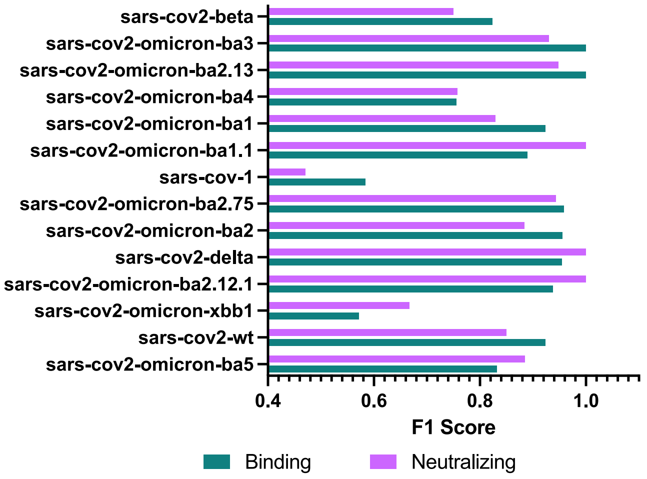


Figure 5. Classification performance of the *ESM2* + *MLM* + *Structure* classifier over all the 14 variants of SARS spike protein (numerical values are available in Supporting Information).

sequence data sets. AbMap fine-tunes foundation models using antibody structure and binding specificity data. Lastly, A2Binder integrates a complex CNN and feed-forward module built upon an LLM and was trained on the CoV-AbDab data set. All these methods were evaluated on held-out test samples from CovAbDab similar to our approach. Table 2 shows that the *ESM2* + *MLM* + *Structure* classifier performed better than all the other approaches with an AUC of 0.9342. Table 2 summarizes the performance comparison of various methods using the area under the Receiver Operating Characteristic curve (ROC AUC), the area under the Precision-Recall curve (PR AUC), and accuracy. We used the metrics for A2Binder from the corresponding paper[23] to ensure a consistent comparison. The trained models was tested on five non-overlapping subsets of the data (5-fold cross-validation); Table 2 reports the average and the standard deviation of all metrics over five independent tests. Observe that our *ESM2+MLM +Struct* achieved the best performance across all metrics.

Structure-Aware LLM Capture More Meaningful Antibody–Antigen Properties. To determine whether our classifiers can provide meaningful latent-space representations of antibody–antigen interactions, we projected the antibody–antigen embeddings produced by the encoder of our classifiers into a 2D space using t-SNE. Figure 6 shows the t-SNE plots for the antibody–antigen embeddings produced by *ESM2* + *MLM* from binding and neutralizing classifiers (Figure 6A,B, respectively), and *ESM2* + *MLM* + *Structure* from binding and neutralizing classifiers (Figure 6C,D, respectively). Points are colored by the binding/neutralizing labels (yellow means not binding or not neutralizing; green means binding or neutralizing). The figure also reports the silhouette score, which measures how similar an object is to its own cluster compared to other clusters. The silhouette score was calculated using the entire embedding vectors. It measures how well data points are clustered, with higher values indicating better separation between groups and more compact clustering within groups. Observe that the structure-aware *ESM2* + *MLM* + *Structure* classifier generates antibody–antigen embeddings which have a stronger separation (i.e., better silhouette score) compared to the sequence-based *ESM2* + *MLM* classifier.

Next, we evaluated the ability of our classifiers to capture the impact of antigenic variations on antibody–antigen interactions. To assess this, (i) we collected the subset of antibodies interacting with all 14 distinct antigen variants, (ii) we produced embeddings for each antibody–antigen pair using the encoder of *ESM2* + *MLM* + *Structure* classifier, (iii) we computed the Euclidean (L_2) distance between the embeddings over all pairs of antigens. We collected these values in a 14 × 14 matrix, where position (i, j) in the matrix represented the average distance between the embedding for antibody A when the input was antigen i and the embedding of A when the input was antigen j. Figure 7A shows the heat map of this matrix, along with a hierarchical clustering of the columns. Observe that antibody–antigen pairs exhibit distinct differences in their embeddings when paired with different antigens. Figure 7B instead illustrates the sequence dissimilarity over all pairs of antigens. Observe that while some of the antigen pairs have highly similar sequences in Figure 7B, the corresponding embeddings are not necessarily similar in Figure 7A. This is because the embeddings have to reflect the functional and structural characteristics of the antibodies interacting with these antigens, which may not be reflected in the sequence similarity. Notably, the embedding differences are significantly larger for SARS-CoV-1, consistent with its distinct functional characteristics compared to SARS-CoV-2.[51] A similar observation can be made for SARS-CoV-2 Omicron XBB.1, which has unique interaction properties with antibodies compared to other SARS-CoV-2 variants.[52,53] Notably, the hierarchical clustering of the embeddings in Figure 7A resembles closely the hierarchical clustering in Figure 7B. SARS-CoV-2 Omicron XBB.1 exhibits high sequence similarity with other variants, positioning itself within the same cluster (red cluster) in Figure 7B. However, despite this sequence similarity, XBB.1 demonstrates a distinct functional response compared to

Table 2. Performance of Dynamic Masking LLM, Ens-Grad, ESM-F, AntiBERTa, AbMap, A2Binder, and our ESM2 + MLM + Structure on the Antibody-Antigen Binding Prediction Problem Using ROC AUC, Precision-Recall AUC and Accuracy[a]

Model	Ref	ROC AUC	PR AUC	Accuracy
Dynamic Masking	55	0.817 (0.004)	0.829 (0.003)	0.737 (0.005)
Ens-Grad	56	0.890 (0.015)	0.859 (0.018)	0.818 (0.018)
ESM-F	23	0.916 (0.012)	0.905 (0.009)	0.839 (0.009)
AntiBERTa2	23	0.918 (0.010)	0.897 (0.014)	0.848 (0.009)
AbMAP	57	0.922 (0.008)	0.908 (0.011)	0.845 (0.014)
A2binder	23	<u>0.930</u> (0.009)	<u>0.923</u> (0.013)	<u>0.861</u> (0.011)
ESM2+MLM+Struct	this study	**0.934** (0.014)	**0.952** (0.017)	**0.874** (0.014)

[a] Each value is reported as mean (standard deviation) (bold numbers indicate the best score, and underlined numbers indicate the second best score).

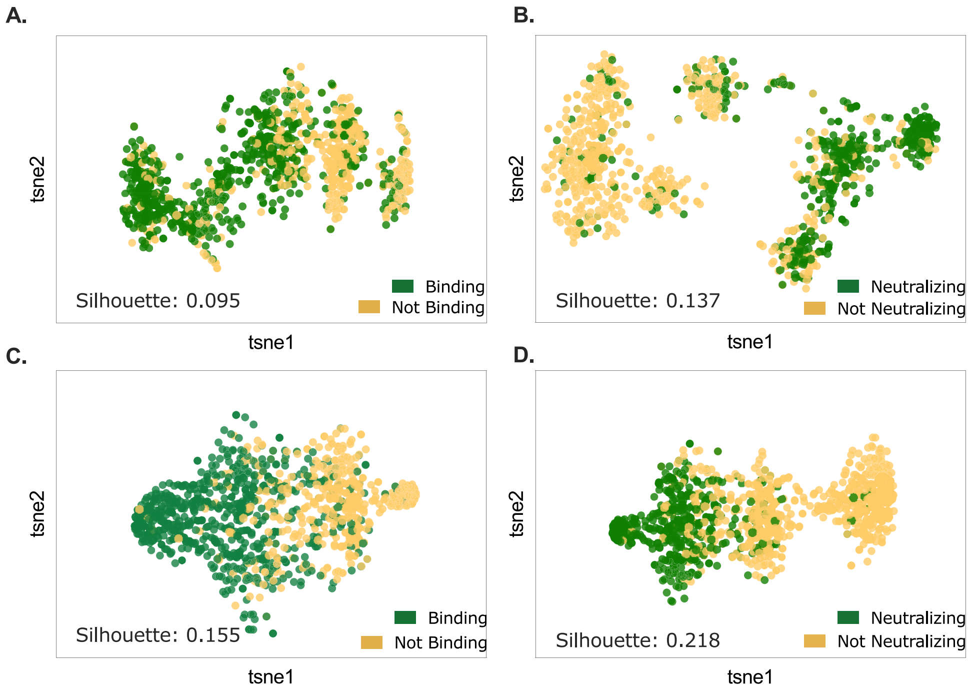


Figure 6. t-SNE 2D projections of the antibody−antigen embeddings produced by *ESM2 + MLM* and *ESM2 + MLM + Structure*; points are colored by the binding/neutralizing labels; silhouette score are reported for clustering; (A) t-SNE projection of the embeddings from the binding classifier *ESM2 + MLM* (finetuned), points colored by binding labels; (B) t-SNE projection of the embeddings from the neutralizing classifier *ESM2 + MLM* (finetuned), points colored by neutralizing labels; (C) t-SNE projection of the embeddings from the binding classifier *ESM2 + MLM + Structure* (finetuned), points colored by binding labels; (D) t-SNE projection of the embeddings from neutralizing classifier *ESM2 + MLM + Structure* (finetuned), points colored by neutralizing labels.

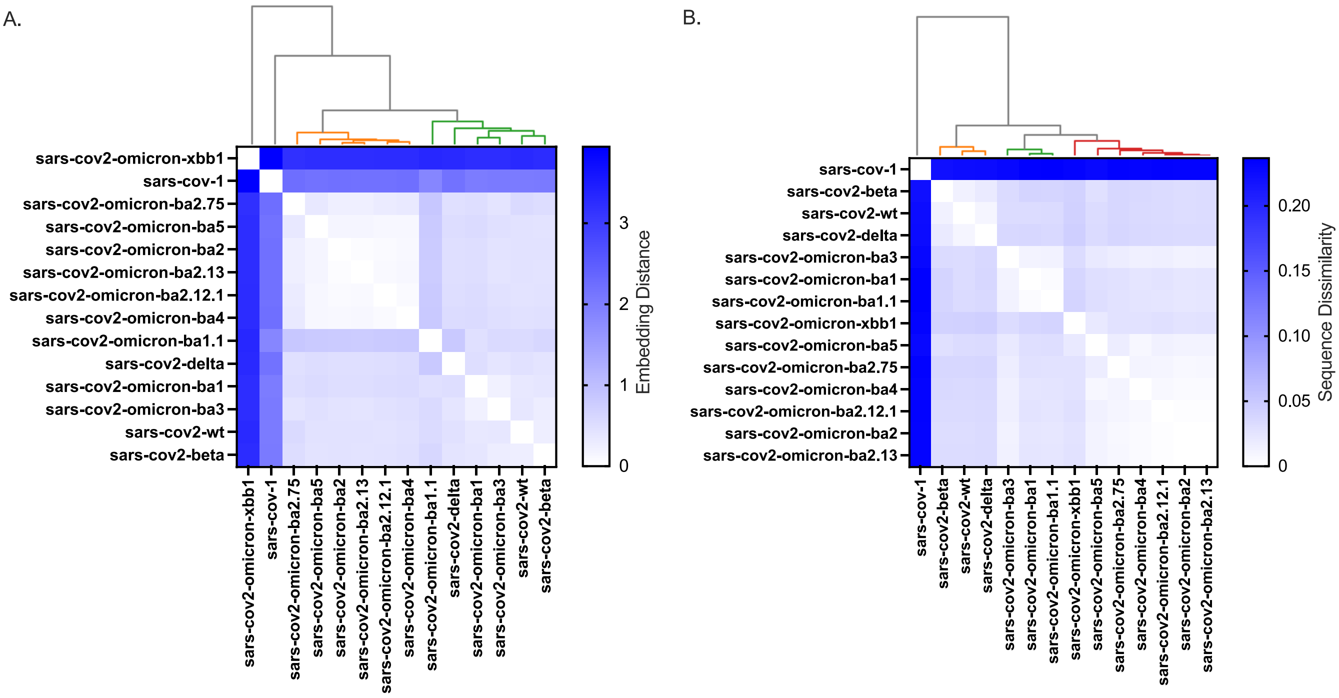


Figure 7. (A) L_2 distance between the embeddings produced by *ESM2 + MLM + Structure* for different choices of the antigens (SARS spike proteins); (B) Sequence dissimilarity between all pairs of SARS spike-proteins variants.

other Omicron BA variants, which is reflected in its separate clustering within the embeddings.

Trained Models Learned to Distinguish Neutralizing Features. Here we dissected the trained models to get some insights on their ability to distinguish between neutralizing and non-neutralizing antibodies that bind to a target antigen. We chose the Omicron-BA-2 variant as the antigen because it had a smallest set of test antibodies (13 neutralizing and 45 non-neutralizing). Our aim was to understand how the models capture neutralizing-specific features, since binding is necessary

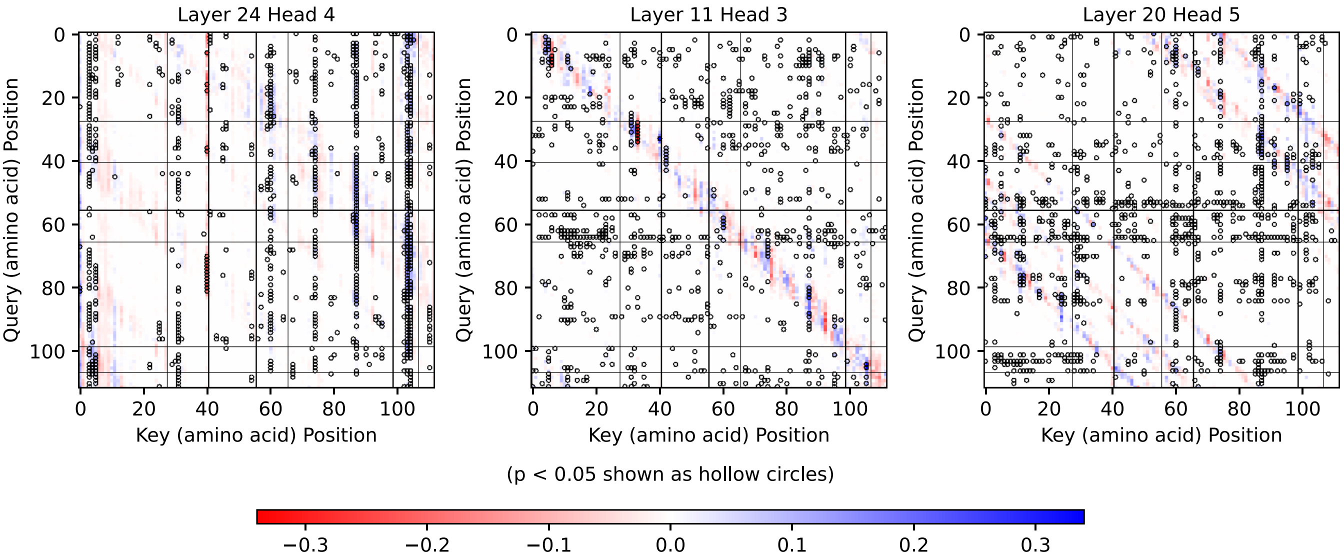


Figure 8. Visualizing the differences in self-attention between neutralizing and non-neutralizing antibodies for the heavy chain classifier; each panel shows the attention difference map (neutralizing minus non-neutralizing) for three most significant attention heads, namely Layer 24 Head 4, Layer 11 Head 3, and Layer 20 Head 5; the *x*-axis indicates key amino acid positions, while the *y*-axis indicates query amino acid positions; red regions indicate higher attention in non-neutralizing antibodies; blue regions indicate higher attention in neutralizing antibodies; hollow circles mark residue pairs with statistically significant differences (pvalue $p < 0.05$); horizontal and vertical lines demarcate the complementarity-determining regions (CDRs) and framework regions (FRs) along the sequence.

but not sufficient for an antibody to neutralize the antigen. It was a natural choice to focus on the attention heads of our architectures. After collecting the values of the attention layers (160 in total) for the heavy chain and the light chain classifiers, we computed the differences in average attention between neutralizing and non-neutralizing antibodies. Figure 8 shows a heat map for the top three most significant attention matrices for the heavy chain classifier. In each heatmap, position (q,k) reports the difference in average attention weight between non-neutralizing and neutralizing antibodies. A red value means the residue at position q has stronger attention to position k in non-neutralizing antibodies, while blue means the opposite. Hollow circles mark (q,k) pairs whose attention differences are statistically significant (p-value $p < 0.05$). Observe that Layer 24 Head 4 shows a strong red signal from FR1 ($q \in [1, 25]$) and FR3 ($q \in [66, 104]$) into CDR1 ($k \in [26, 38]$), suggesting that non-neutralizing antibodies anchor scaffold residues onto CDR1. In contrast, blue shading across CDR3 keys ($k \in [96, 107]$) indicates that neutralizing antibodies broadly target CDR3. Layer 11 Head 3 displays alternating red and blue patches along the main diagonal ($q \approx k$), reflecting class-specific modulation of local, neighbor-level attention. Layer 20 Head 5 reveals a long-range blue signal connecting CDR1 queries ($q \in [26, 38]$) to FR3/CDR3 keys ($k \in [80, 111]$), highlighting a neutralizer-specific dependency between CDR and the framework regions. Together, these patterns show that the model differentiates neutralizing from non-neutralizing antibodies through distinct short- and long-range interactions: non-neutralizers rely on CDR1-based scaffolding, whereas neutralizers center CDR3 as a key hub for high-affinity binding. We also observed significant differences between these two groups of antibodies in the light chain classifier. All the differences in attention heads are presented in the Supporting Information.

■ DISCUSSION

In this work, we studied whether it is advantageous to incorporate the 3D structure of antibodies into LLMs, making them "structure-aware" while retaining their ability to learn the "protein language". We evaluated the performance of our structure-aware LLMs against traditional sequence-based LLMs. We also compared the performance of antibody—antigen interaction classifiers using task-specific finetuned versions of these LLMs. To the best of our knowledge, we are the first to address the problem of predicting neutralizing properties of antibodies against the SARS-CoV-2 spike protein.

Our findings demonstrate the superior performance of structure-aware LLMs in antibody—antigen interaction predictions, in particular when finetuned from the pretrained protein language model ESM2. We showed that while ESM2 was designed as a foundational model to capture the general "protein language", it can be significantly improved for antibody—antigen interaction prediction by refining it using target specific antibody sequences and structures. In the fine-tuned configuration, the AUC for the binding prediction task improved from 0.8417 (*ESM2* + *MLM*) to 0.9342 (*ESM2* + *MLM* + *Structure*). Similarly, for the neutralizing prediction task, the AUC increased from 0.8757 (*ESM2* + *MLM*) to 0.9538 (*ESM2* + *MLM* + *Structure*). Our results indicate that the finetuned configuration consistently outperforms the frozen configuration, demonstrating the importance of allowing antibody-specific LLMs to adapt their parameters for improved prediction accuracy in binding and neutralization tasks. Finetuning enables the model parameters to specialize, capturing detailed, task-specific features critical for antibody—antigen interactions. Conversely, the frozen configuration relies solely on general, pretrained features, limiting the model's ability to adapt to nuanced structural and functional characteristics necessary for accurate predictions.

Our antibody-specific LLMs consistently outperformed both pretrained and fine-tuned ESM2 on binary classification tasks (binding or neutralization), emphasizing their strong capability to capture antibody-specific functional patterns. However, for the multiclass prediction tasks (bind + neutralize, bind + no neutralize, neither) the finetuned ESM2 slightly outperformed our models. This is likely due to the fact that the finetuned ESM2 model has a broader knowledge of protein sequence diversity and enables it to identify subtle structural or evolutionary differences critical for nuanced multiclass predictions, which our specialized models may not fully capture due to their narrower antibody-focused training. Nevertheless, our ensemble model which combines the binding and neutralizing prediction classifiers outperforms ESM2 in this task.

Our *ESM2 + MLM + Structure* classifier outperformed DeepAIR, Dynamic Masking LLM, Ens-Grad, ESM-F, AntiBERTa, AbMap, and A2Binder on the binding prediction problem. The key advantage of our approach is to incorporate the structural information by training the model to predict the contact map. This structure-aware learning enhances binding and neutralization prediction, as antibody functionality is inherently linked to its 3D structure. Although A2binder's performance is very close to ours, our classifier is much simpler. We use only a basic feed-forward layer for classification, whereas A2Binder incorporates a more complex CNN and feed-forward module on top of the LLM. Pretrained LLMs on antibody sequence data sets, such as ESM-F and AntiBERTa, show similar performance. However, the superior performance of A2Binder and ESM2 + MLM + Struct indicates that pretraining LLMs on target-specific antibody data sets enhances downstream task performance. We could not find any published tool for predicting neutralizing antibodies. We also showed that the *ESM2 + MLM + Structure* classifier generates more meaningful antibody−antigen embeddings compared to the *ESM2 + MLM* classifier.

In the absence of experimentally determined antibody 3D structures, we used ESMFold to generate structural predictions. These predicted structures were used to enhance the antibody language model. While ESMFold captures overall structural topology with high efficiency, it lacks the atomic-level precision of experimental methods such as X-ray crystallography or cryo-EM. Despite these limitations, incorporating ESMFold-derived structural features led to consistent improvements in downstream model performance. This suggests that even approximate structural context provides meaningful biophysical cues, making the approach both effective and scalable in scenarios where experimental structures are unavailable.

Despite these encouraging results, there is still room for improvement. A larger training data set is likely to further enhance model performance by improving its ability to generalize across diverse antibody−antigen interactions. Additionally, integrating residue-level details, such as solvent accessibility, secondary structure, and protein−protein binding interfaces, could provide a more comprehensive representation of functional determinants. Features like proximity to key binding sites, including the receptor-binding site and the furin-binding site, may also improve the predictive accuracy. Future work could explore these directions to refine antibody-specific LLMs and extend their applications to broader antibody−antigen interaction studies.

■ ASSOCIATED CONTENT

Data Availability Statement

All the data sets, the code, and the weights for the models described in this study are publicly available at https://github.com/ucrbioinfo/Structure-Aware-LLM-Ab-Ag-Interaction.git.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c00973.

Tables with numeric values for the barplot figures, additional visualizing figures (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Stefano Lonardi − *Department of Computer Science and Engineering, University of California, Riverside, California 92521, United States;* orcid.org/0000-0002-2696-7274; Email: stelo@ucr.edu

Authors

Faisal Bin Ashraf − *Department of Computer Science and Engineering, University of California, Riverside, California 92521, United States;* orcid.org/0000-0003-4006-5389

Vinz Angelo Madrigal − *Department of Computer Science and Engineering, University of California, Riverside, California 92521, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c00973

Author Contributions

F.B.A.: conceived the project, designed the methodological approach, implemented the model, performed the analyses, collected experimental results, and drafted the manuscript. V.A.M.: was responsible for data collection, preprocessing, and data cleaning. S.L.: provided financial support, supervised the study and edited the manuscript. All authors contributed to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Janeway, C. A.; Travers, P.; Walport, M.; Shlomchik, M. *Immunobiology: The Immune System in Health and Disease*; Garland Science, 2001.

(2) Kabat, E. A.; Wu, T. T. Attempts to locate complementarity-determining residues in the variable positions of light and heavy chains. *Ann. N. Y. Acad. Sci.* **1971**, *190*, 382−393.

(3) Davies, D. R.; Metzger, H. Structural basis of antibody function. *Annu. Rev. Immunol.* **1983**, *1*, 87−117.

(4) Padlan, E. A. Anatomy of the antibody molecule. *Mol. Immunol.* **1994**, *31*, 169−217.

(5) Honegger, A.; Plückthun, A. Yet another numbering scheme for immunoglobulin domains. *J. Mol. Biol.* **2001**, *309*, 657−670.

(6) Wedemayer, G. J.; Patten, P. A.; Wang, L. H.; Schultz, P. G.; Stevens, R. C. Structural insights into the evolution of an antibody combining site. *Science* **1997**, *276*, 1665−1669.

(7) Björkman, P. J.; Saper, M. A.; Samraoui, B.; Bennett, W. S.; Strominger, J. L.; Wiley, D. C. Structure of the human class I histocompatibility antigen, HLA-A2. *Nature* **1987**, *329*, 506−512.

(8) Xu, R.; Ekiert, D. C.; Krause, J. C.; Hai, R.; Crowe, J. E.; Wilson, I. A. Structural basis of preexisting immunity to the 2009 H1N1 pandemic influenza virus. *Science* **2013**, *340*, 1339−1345.

(9) AlQuraishi, M. Machine learning in protein structural modeling. *Curr. Opin. Chem. Biol.* **2019**, *65*, 126−134.

(10) Kim, J.; McFee, M.; Fang, Q.; Abdin, O.; Kim, P. M. Computational and artificial intelligence-based methods for antibody development. *Trends Pharm. Sci.* **2023**, *44*, 175−189.

(11) Ruffolo, J. A.; Sulam, J.; Gray, J. J. Antibody structure prediction using interpretable deep learning. *Patterns* **2022**, *3*, No. 100406, DOI: 10.1016/j.patter.2021.100406.

(12) Li, M.; Shi, Y.; Hu, S.; Hu, S.; Guo, P.; Wan, W.; Zhang, L. Y.; Pan, S.; Li, J.; Sun, L.; et al. MVSF-AB: Accurate antibody-antigen binding affinity prediction via multi-view sequence feature learning. *Bioinformatics* **2024**, *41*, No. btae579.

(13) Xu, Z.; Davila, A.; Wilamowski, J.; Teraguchi, S.; Standley, D. M. Improved antibody-specific epitope prediction using alphafold and abadapt. *ChemBioChem* **2022**, *23*, No. e202200303.

(14) Viswanathan, R.; Carroll, M.; Roffe, A.; Fajardo, J. E.; Fiser, A. Computational prediction of multiple antigen epitopes. *Bioinformatics* **2024**, *40*, No. btae556.

(15) Huang, Y.; Zhang, Z.; Zhou, Y. AbAgIntPre: A deep learning method for predicting antibody-antigen interactions based on sequence information. *Front. Immunol.* **2022**, *13*, No. 1053617.

(16) Mason, D. M.; Friedensohn, S.; Weber, C. R.; Jordi, C.; Wagner, B.; Meng, S. M.; Ehling, R. A.; Bonati, L.; Dahinden, J.; Gainza, P.; et al. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat. Biomed. Eng.* **2021**, *5*, 600−612.

(17) Qiu, T.; Zhang, L.; Chen, Z.; Wang, Y.; Mao, T.; Wang, C.; Cun, Y.; Zheng, G.; Yan, D.; Zhou, M.; et al. SEPPA-mAb: spatial epitope prediction of protein antigens for mAbs. *Nucleic Acids Res.* **2023**, *51*, W528−W534.

(18) Curion, F.; Theis, F. J. Machine learning integrative approaches to advance computational immunology. *Genome Med.* **2024**, *16*, No. 80, DOI: 10.1186/s13073-024-01350-3.

(19) Hie, B. L.; Shanker, V. R.; Xu, D.; Bruun, T. U.; Weidenbacher, P. A.; Tang, S.; Wu, W.; Pak, J. E.; Kim, P. S. Efficient evolution of human antibodies from general protein language models. *Nat. Biotechnol.* **2024**, *42*, 275−283.

(20) Sun, T.; Zhou, B.; Lai, L.; Pei, J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinf.* **2017**, *18*, No. 277.

(21) Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7112−7127.

(22) Chakraborty, C.; Bhattacharya, M.; Pal, S.; Lee, S.-S. Prompt engineering-enabled llm or mllm and instigative bioinformatics pave the way to identify and characterize the significant sars-cov-2 antibody escape mutations. *Int. J. Biol. Macromol.* **2025**, *287*, No. 138547.

(23) He, H.; He, B.; Guan, L.; Zhao, Y.; Jiang, F.; Chen, G.; Zhu, Q.; Chen, C. Y.-C.; Li, T.; Yao, J. De novo generation of SARS-CoV-2 antibody CDRH3 with a pre-trained generative large language model. *Nat. Commun.* **2024**, *15*, No. 6867.

(24) Casadio, R.; Martelli, P. L.; Savojardo, C. Machine learning solutions for predicting protein-protein interactions. *WIREs Comput. Mol. Sci.* **2022**, *12*, No. e1618.

(25) Luo, Y.; Jiang, G.; Yu, T.; Liu, Y.; Vo, L.; Ding, H.; Su, Y.; Qian, W. W.; Zhao, H.; Peng, J. ECNet is an evolutionary context-integrated deep learning framework for protein engineering. *Nat. Commun.* **2021**, *12*, No. 5743.

(26) AlQuraishi, M. Machine learning in protein structure prediction. *Curr. Opin. Chem. Biol.* **2021**, *65*, 1−8.

(27) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.

(28) Alexov, E.; Honig, B. *Handbook of Cell Signaling*; Elsevier, 2010; pp 11−13.

(29) Schelhorn, S.-E.; Lengauer, T.; Albrecht, M. An integrative approach for predicting interactions of protein regions. *Bioinformatics* **2008**, *24*, i35−i41.

(30) Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **2022**, *50*, D439−D444.

(31) Ali, S.; Chourasia, P.; Patterson, M. When Protein Structure Embedding Meets Large Language Models. *Genes* **2024**, *15*, 25.

(32) Wang, Y.; Ribeiro, J. M. L.; Tiwary, P. Machine learning approaches for analyzing and enhancing molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2020**, *61*, 139−145.

(33) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123−1130.

(34) Polonsky, K.; Pupko, T.; Freund, N. T. Evaluation of the ability of AlphaFold to predict the three-dimensional structures of antibodies and epitopes. *J. Immunol.* **2023**, *211*, 1578−1588.

(35) Rao, R. M.; Liu, J.; Verkuil, R.; Meier, J.; Canny, J.; Abbeel, P.; Sercu, T.; Rives, A.*MSA transformer*; Proceedings of the 38th International Conference on Machine Learning; PMLR, 2021; pp 8844−8856.

(36) Wang, D.; Pourmirzaei, M.; Abbas, U. L.; Zeng, S.; Manshour, N.; Esmaili, F.; Poudel, B.; Jiang, Y.; Shao, Q.; Chen, J.; et al. S-PLM: Structure-aware Protein Language Model via Contrastive Learning between Sequence and Structure. *Adv. Sci.* **2023**, *12*, No. 2404212.

(37) Chen, C. S.; Zhou, J.; Wang, F.; Liu, X.; Dou, D. Structure-aware protein self-supervised learning. *Bioinformatics* **2023**, *39*, No. btad189.

(38) Su, J.; Han, C.; Zhou, Y.; Shan, J.; Zhou, X.; Yuan, F.Saprot: Protein language modeling with structure-aware vocabulary *bioRxiv* 2023 .

(39) Zhang, Z.; Lu, J.; Chenthamarakshan, V.; Lozano, A.; Das, P.; Tang, J. Structure-informed protein language model. *arXiv preprint arXiv:2402.05856* 2024,.

(40) Sela-Culang, I.; Kunik, V.; Ofran, Y. The structural basis of antibody-antigen recognition. *Front. Immunol.* **2013**, *4*, No. 302.

(41) Raybould, M. I. J.; Kovaltsuk, A.; Marks, C.; Deane, C. M. CoV-AbDab: the coronavirus antibody database. *Bioinformatics* **2021**, *37*, 734−735.

(42) Rose, P. W.; Bi, C.; Bluhm, W. F.; Christie, C. H.; Dimitropoulos, D.; Dutta, S.; Green, R. K.; Goodsell, D. S.; Prlić, A.; Quesada, M.; et al. The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res.* **2012**, *41*, D475−D482.

(43) Ezkurdia, I.; Grana, O.; Izarzugaza, J. M.; Tress, M. L. Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. *Proteins:Struct., Funct., Bioinf.* **2009**, *77*, 196−209.

(44) Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. arXiv:1810.04805. arXiv.org e-Printarchive. https://doi.org/10.48550/arXiv.1810.04805.

(45) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.et al.Language models of protein sequences at the scale of evolution enable accurate structure prediction *bioRxiv* 2022 DOI: 10.1101/2022.07.20.500902.

(46) Nahali, S.; Safari, L.; Khanteymoori, A.; Huang, J. StructmRNA a BERT based model with dual level and conditional masking for mRNA representation. *Sci. Rep.* **2024**, *14*, No. 26043.

(47) Iuchi, H.; Matsutani, T.; Yamada, K.; Iwano, N.; Sumi, S.; Hosoda, S.; Zhao, S.; Fukunaga, T.; Hamada, M. Representation learning applications in biological sequence analysis. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 3198−3208.

(48) Chu, S. K. S.; Narang, K.; Siegel, J. B. Protein stability prediction by fine-tuning a protein language model on a mega-scale dataset. *PLoS Comput. Biol.* **2024**, *20*, No. e1012248.

(49) Kulmanov, M.; Guzmán-Vega, F. J.; Roggli, P. D.; Lane, L.; Arold, S. T.; Hoehndorf, R. Protein function prediction as approximate semantic entailment. *Nat. Mach. Intell.* **2024**, *6*, 220−228.

(50) Schmirler, R.; Heinzinger, M.; Rost, B. Fine-tuning protein language models boosts predictions across diverse tasks. *Nat. Commun.* **2024**, *15*, No. 7407.

(51) Zhang, Z.-B.; Xia, Y.-L.; Shen, J.-X.; Du, W.-W.; Fu, Y.-X.; Liu, S.-Q. Mechanistic origin of different binding affinities of SARS-CoV and SARS-CoV-2 spike RBDs to human ACE2. *Cells* **2022**, *11*, No. 1274.

(52) Qu, P.; Faraone, J. N.; Evans, J. P.; Zheng, Y.-M.; Carlin, C.; Anghelina, M.; Stevens, P.; Fernandez, S.; Jones, D.; Panchal, A. R.; et al. Enhanced evasion of neutralizing antibody response by Omicron XBB. 1.5, CH. 1.1, and CA. 3.1 variants. *Cell Rep.* **2023**, *42*, No. 112443, DOI: 10.1016/j.celrep.2023.112443.

(53) Sharma, T.; Gerstman, B.; Chapagain, P. Distinctive features of the XBB. 1.5 and XBB. 1.16 spike protein receptor-binding domains and their roles in conformational changes and angiotensin-converting enzyme 2 binding. *Int. J. Mol. Sci.* **2023**, *24*, No. 12586.

(54) Zhao, Y.; He, B.; Xu, F.; Li, C.; Xu, Z.; Su, X.; He, H.; Huang, Y.; Rossjohn, J.; Song, J.; Yao, J. DeepAIR: A deep learning framework for effective integration of sequence and 3D structure to enable adaptive immune receptor analysis. *Sci. Adv.* **2023**, *9*, No. eabo5128.

(55) Ng, K.; Briney, B. Focused learning by antibody language models using preferential masking of non-templated regions. *Patterns* **2025**, *6*, No. 101239, DOI: 10.1016/j.patter.2025.101239.

(56) Liu, G.; Zeng, H.; Mueller, J.; Carter, B.; Wang, Z.; Schilz, J.; Horny, G.; Birnbaum, M. E.; Ewert, S.; Gifford, D. K. Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics* **2020**, *36*, 2126−2133.

(57) Singh, R.; Im, C.; Qiu, Y.; Mackness, B.; Gupta, A.; Joren, T.; Sledzieski, S.; Erlach, L.; Wendt, M.; Nanfack, Y. F.; et al. Learning the language of antibody hypervariability. *Proc. Natl. Acad. Sci. U.S.A.* **2025**, *122*, No. e2418918121.

CAS BIOFINDER DISCOVERY PLATFORM™
PRECISION DATA FOR FASTER DRUG DISCOVERY
CAS BioFinder helps you identify targets, biomarkers, and pathways
Unlock insights
CAS
A division of the American Chemical Society